Reconsidering the Smart Metering Data Collection Frequency for Distribution State Estimation

Qipeng Chen, Dritan Kaleshi
and Simon Armour
Communication Systems and Networks Group
Department of Electrical and Electronic Engineering
University of Bristol,
United Kingdom
{Qipeng.Chen, Dritan.Kaleshi, Simon.Armour}@bristol.ac.uk

Zhong Fan
Toshiba Research Europe Limited,
Telecommunications Research Laboratory,
Bristol, United Kingdom
Zhong.Fan@toshiba-trel.com

***Abstract*—The current UK Smart Metering Technical Specification requires smart meter readings to be collected once a day, primarily to support accurate billing without violating users' privacy. In this paper we consider the use of Smart Metering data for Distribution State Estimation (DSE), and compare the effectiveness of daily data collection strategy with a more frequent, half-hourly SM data collection strategy. We first assess the suitability of using the data for load forecasting at Low Voltage (LV) transformers, and then use the forecast for DSE. The outputs of DSE indicate a whole system's real-time status which can be used to make effective decisions for grid control. Our statistical test results show that the use of the half-hourly collected SM data significantly improves the load forecasting accuracy. However, the DSE results show that neither data collection strategy alone is sufficient to estimate a system's status accurately, and both require additional real-time measurements, with significantly fewer additional measurements points required if the data is collected half-hourly. This research offers a unique DSE perspective which will provide evidence towards a more comprehensive specification of the SM data collection frequency if it is to be used for smart grid operational support.**

I. Introduction

Smart metering (SM) can provide raw measurement data at the edge of the electrical power system to support smart grid functions in addition to using the data for more and timely accurate billing. In the past few years there is significant interest to explore the potential *bonus* of the data usage for power system control, but the lack of available linked data to use for this purpose has hampered some of this activity. The current UK Smart Metering Technical Specification (SMETS) specifies that smart meter readings taken on half-hourly intervals are collected on the enterprise side only once a day. This strategy has been primarily proposed to support accurate billing purposes without violating users' privacy. Nevertheless, as described in [1], more frequent SM data collection is technically feasible. In this paper, we also consider the case of half-hourly SM data collection where the meter readings of time $t-1$ are collected at t, where the duration in between is 0.5 hour.

Distribution system control requires measurement of several electrical system variables (hereafter referred to as *measurements*) at Low Voltage (LV) transformers, in especially their voltages. However, the transformers are rarely equipped with sensors due to the high costs [2]. Before the SM system was introduced, the transformers' loads were modeled monthly or for even longer periods [3]. Given all the transformers' loads and other available measurements in a power circuit, their voltages are estimated by Distribution State Estimation (DSE). DSE filters out the errors from the provided data, and estimates voltage magnitudes and phasor angles of power system points/nodes where the measurements are not directly provided [4]. Using the model-generated loads as inputs to estimate voltages is reliable only when the system is stable during the observation and computation window. The introduction of distributed and renewable generation, and electrical vehicles, and novel grid applications such as Demand Side Management will make loads more dynamic within the DSE window, introducing difficulties in system control. The use of SM data offers a new opportunity, at no additional cost for the power system operators, for load modeling. A transformer's loads can be directly represented by the aggregation of the smart meter load measurements for all the customers downstream from a given LV transformer. The loads are then used as inputs to the same DSE in order to obtain an estimation of the system state (i.e. power system transformer node voltages).

The use of SM data for DSE has been reported in the literature. [5] shows that DSE based on the SM data achieves more accurate outputs than without the SM data, as expected. A robust state estimator combined with a machine learning algorithm is proposed in [2] as a close-loop procedure for using SM data for DSE. This overcomes the issue when some of the smart meters cannot transmit data in real-time. Both these works above assume that the SM data is collected immediately after measuring, i.e. real-time SM data is available. However, as specified by the UK SM system [6], the smart meters measure half-hourly, and the half-hourly readings are collected at the end of the day. Therefore, the latest SM data available to use is from the previous day. [7] uses the daily collected SM data for DSE. It directly represents the transformer's loads with the previous day's data, and it achieves good accuracy when additional real-time measurements are provided.

So far there is no clear comparison on the effect of the SM data collection frequencies on the performance (accuracy) of DSE. We compare the performance of DSE with SM data collection frequencies of one day and half hour respectively. Similar to the method in [7], we use another naive model under the assumption of half-hourly SM data collection; the load from the previous half hour is taken to represent the real-time load of the transformer and then taken on to proceed with DSE. Our research findings can be used to encourage grid operators and practitioners to determine the SM data collection frequency with more comprehensive considerations, not only from the perspectives of billing function and user privacy, but also from the perspective of supporting power system control.

To go beyond the base comparison to previous works, we also introduce one state of the art load forecasting technique to forecast the loads at the LV transformers, which can achieve better accuracy compared with the naive methods. Load forecasting is a very well researched area [8]. Artificial Neural Network (ANN) is one of the most outstanding nonlinear learners. It has been used in [3], [9] to forecast large areas' loads of the next day. The method in [10] also uses ANN as the learner to forecast the load at the next time point for a small area. However, several drawbacks limit the use of ANN in real world applications. First, it easily gets stuck in describing the random errors and noises instead of deriving the relation between the dependent and independent variables, which is called *over fitting*. Second, ANN performs poorly when handling high dimensional data because the computation cost increases exponentially with the increase of the data dimensions. Moreover, as described in [11], the time series loads are linearly auto-correlated, so ANN does not stand a better chance than a simpler linear model in load forecasting. Support Vector Regression (SVR) is one of the comparable techniques for load forecasting. It handles either linear or nonlinear related variables by switching the kernel functions. In [12] it is shown that SVR overcomes the above issues of ANN and a framework with 24 SVR models is used to forecast the loads at different hours of the day separately. Utilizing multiple models not only reduces the computation costs but also increases the load forecasting accuracy. Different versions of SVR for load forecasting are reported in [13] and [14]. We use the framework in [12] for LV transformer load forecasting. It was originally designed with SVR as the learner. We adapt ANN to the framework as well. Most of the papers above concern load forecasting for a large area, where the exogenous variables including the weather and temperature show high relevance with the loads. Therefore these variables are usually involved in forecasting. However, as described in [15] these variables have few correlations with the loads of a small area. Therefore in this paper we only consider the historical loads as the inputs of load forecasting.

This paper is organized as follows. The DSE problem with smart meter data is specified in Section II. The load forecasting techniques and the DSE method are briefly described in Section III. The simulation results are summarized and discussed in Section IV. Conclusions are drawn in Section V.

II. Problem Specification

We use the 33 node Medium Voltage (MV) power network in [16] as an example to explain how the SM data can facilitate system control. As shown in Fig. 1, the circuit interconnects one Grid Supply Point (GSP) (node 1) and 32 LV transformers (node 2-33). The GSP is the only node equipped with sensors, so only the loads and voltages at this node are observable. The same information is required at all of the transformers (especially their voltages) in order to obtain the system's operation condition. Therefore we use DSE to estimate the LV transformers' voltages with the SM data.

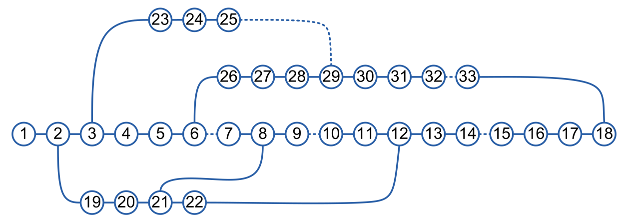

Fig. 1. The 33 node MV network.

Each of the transformers is the starting point of an LV circuit which consists of around 150 households. We assume the households have already been equipped with smart meters. A meter measures both active power and reactive power in every half an hour. The active power P_t^b of transformer b at time t is approximately equal to the sum of the active power of all the households downstream from b, and is given by:

$$P_t^b \approx \sum\nolimits_{n=1}^{S_n} p_{t,n}^b, \tag{1}$$

where S_n is the total number of households on the circuit, n is a household's index, and p is the smart meter measured active power. The reactive power Q_t^b is calculated in the same way. The power losses along transmission are not taken into account as these are much smaller than the loads.

However the SM data is not immediately collected after measurement. For the daily SM data collection the latest meter readings are from the day before in 48 half-hourly segments. Therefore, the latest loads of b at hand at time t are P_{t-48}^b and Q_{t-48}^b. The delay of the SM data transmission of the half-hourly SM data collection strategy is 30 minutes, so the latest loads of b at t are P_{t-1}^b and Q_{t-1}^b. In [7] a naive model is used to represent the real-time loads of the transformer when smart meter readings are collected once a day. The transformer's real-time loads are represented by the loads of this transformer at the same time of the previous day. We propose another naive model which is applicable to the half-hourly SM data collection strategy: we use a transformer's loads of one half hour before to represent the transformer's loads at the real time; in other words, we assume that the power load at transformer remains, averagely, constant over the half hour interval. The two models are given by:

- *Naive Model 1*: $P_t^b \approx P_{t-48}^b \quad Q_t^b \approx Q_{t-48}^b$ (from [7])
- *Naive Model 2*: $P_t^b \approx P_{t-1}^b \quad Q_t^b \approx Q_{t-1}^b$

where the duration between $t-1$ and t is 0.5 hour.

Load forecasting serves the same purpose as the above two methods, but it provides more accurate estimates of the loads of the transformer using both the information of the past loads and possibly other exogenous data. The forecasting of the active power for the daily and half-hourly SM data collection strategies are given by Eq.2a and 2b respectively.

$$P_t^b \approx E(x_t^1, x_t^2, x_t^3, \ldots, P_{t-48}^b, P_{t-49}^b, \ldots, P_{t-i}^b, \ldots) \tag{2a}$$

$$P_t^b \approx F(x_t^1, x_t^2, x_t^3, \ldots, P_{t-1}^b, P_{t-2}^b, \ldots, P_{t-j}^b, \ldots) \tag{2b}$$

P^b_{t-i} is a *load variable* which denotes the active power of the transformer i half an hour prior to t. i and j are no smaller than 48 and 1 respectively. $\boldsymbol{x_t}$ is the set of *exogenous variables*, such as the indexes of the day type (Monday, Tuesday, $\ldots$, Sunday) and the month of t. If this is the load forecasting problem for a large area, $\boldsymbol{x_t}$ should further cover the weather and temperature at the time of forecast. However, we aim to forecast the loads at the LV transformers, where the temperature and weather show less relevance with the load demands. The set of load and exogenous variables are called attributes in the area of machine learning. P^b_t is called the *target*. The combination of the target and all attributes values of a specific time is called an $instance$ which is usually in the form of a vector. Let's say the SM data is collected daily and the time is τ. The corresponding instance is represented as $\{x^1_\tau,\, x^2_\tau,\, x^3_\tau,\, \ldots,\, P^b_{\tau-48},\, P^b_{\tau-49},\, \ldots,\, P^b_{\tau-i},\, \ldots,\, \boldsymbol{P^b_\tau}\}$. The last element of the instance is the target value. E and F are the unknown functions which denote the relations between the attributes and the target for the two frequencies of SM data collection.

Load forecasting mainly consists of four steps. In the first step, relevant attributes are selected. The exogenous variables are usually determined by experience, but the selection of the load variables is more difficult. The most straightforward way is to use the full set of the load variables. For example, we can use the previous two weeks' loads, $\{P^b_{t-1},\, P^b_{t-2},\, \ldots,\, P^b_{t-672}\}$ to forecast P^b_t when the data is collected half-hourly. However, the use of the full set can easily cause over fitting. Therefore, a more concise set should be used. For example, one can choose $\{P^b_{t-48}, P^b_{t-336}\}$ as the load variables when the SM data is collected once a day, because the analysis results show the loads from the day before and the week before have high correlations with P^b_t. In the second step, a training set is generated from the available historical data. The training set is usually in the form of a matrix, where a row is an instance. For example, in order to forecast P^b_τ, we use the past 200 days' loads to generate the training set. The corresponding training set could consist of the instances whose target values are $P^b_{\tau-9600}$, $P^b_{\tau-9599}$, $P^b_{\tau-9598}$, $\ldots$, and $P^b_{\tau-48}$ respectively. Note that the training size has a significant impact on load forecasting. A big training set easily causes over fitting, and a small one could lead to information losses. In the third step, a proper *learner* is trained by the training set. The selection of the learner depends on the relation between the target and the attributes. In this work we use both ANN and linear-SVR as the learners which handle the nonlinear and linear relations respectively. In the final step, the well trained learner forecasts the target load value with the latest attribute values. The reactive power Q^b_t is forecasted in the same way.

Once the active and reactive powers of all 32 LV transformers at time t are obtained either by a naive model or a load forecasting technique, we apply DSE for the MV system, which infers the transformers' voltages at that same time.

III. Methodology

We use the strategic, seasonality-adjusted support vector regression machines based model (SSA-SVR) in [12] for load forecasting. This approach is applicable to both frequencies of SM data collection. SVR is used as the learner with the radial basis function (RBF) as the kernel function to handle the nonlinear data relation. In this research we switch the kernel function which leads to the linear-SVR, and we also use ANN as the learner. This allows us to look at the differences between ANN and SVR, and the differences between the linear and nonlinear learners for load forecasting under the model. We use the Weighted Least Square (WLS) state estimator in [4] for DSE. Details of the two algorithms are given in the following subsections.

A. The Load Forecasting Approach

48 parallel models are used, where each model forecasts the loads of a unique half hour of the day. These models work in the same way but separately. Therefore, in the following we explain the implementation of a single model only.

Firstly, the attributes of the model are chosen. The exogenous variables are the indices of the day type (Monday, Tuesday, $\ldots$, Sunday), the month and the holidays. All of the indices are in the form of binary variables. We use the first 7 and 12 attributes to describe the day type and the month respectively. For example, the values of these attributes are $\{\overline{1,0,0,0,0,0,0},\ \overline{1,0,0,0,0,0,0,0,0,0,0,0}\}$ if the instance is from a Monday in January. The next 4 attributes are used to denote if the current, previous and the next two days are holidays (1 is true, 0 is false). The load variables are selected by the feature selection mechanisms. Our principle of feature selection is to choose a subset of load variables which have high correlations with the target, yet have low inter-correlations. The load variables are selected from the lagged loads of the two weeks prior to the time of the target variable. For the daily or half-hourly SM data collection, the load variables are chosen from $\{P^b_{t-48}, P^b_{t-49}, \ldots, P^b_{t-672}\}$ or $\{P^b_{t-1},\, P^b_{t-2},\, \ldots,\, P^b_{t-672}\}$ respectively.

Secondly, the training set is prepared. The model is specialized for a single time of the day, so the training set only includes the instances of the same time point. The loads of the past whole year are used to generate the training set. In order to forecast P^b_τ, the target values of the instances in the training set are $P^b_{\tau-48\times365}$, $P^b_{\tau-48\times364}$, $P^b_{\tau-48\times363}$, $\ldots$ till $P^b_{\tau-48}$. The attributes have been selected in the first step. Their values are determined considering the target values. Thirdly, we use either linear-SVR or ANN as the learner. Hereafter the technique using linear-SVR or ANN as the learner is simply denoted by *SSA-SVR* or *SSA-ANN*. There is no parameter which needs to be pre-defined for linear-SVR. However, the performance of ANN highly depends on the number of the hidden neurons. In this research, the number of the hidden neurons for each separate model is determined by the trial-and-error method. Finally, the 48 learners are well trained by the corresponding training sets. Each learner forecasts the loads at a single time point of the day. The approach is used for the 32 LV transformers in Fig. 1. Therefore, both the real-time active and reactive powers at the transformers can be forecasted.

B. The WLS DSE Algorithm

Once the active and reactive powers at all LV transformers are forecasted, we use the WLS state estimator for the MV system. The state estimator aims to obtain the states of the MV system which minimize the overall errors involved in all measurements. This is an optimization problem and is formulated as:

$$\begin{aligned}\min_{\boldsymbol{x}} j(\boldsymbol{x}) &= \sum\nolimits_{i=1}^{Z_n}(z_i - h_i(\boldsymbol{x}))^2/R_{ii} \\ &= [\boldsymbol{z} - h(\boldsymbol{x})]^T R^{-1}[\boldsymbol{z} - h(\boldsymbol{x})]\end{aligned} \tag{3}$$

where $\boldsymbol{x}$ is the state of the system including the phasor angles (θ) and voltage magnitudes (V) of all of the nodes. $\boldsymbol{x}$ at time t is represented as the vector of [θ_t^2, ..., θ_t^i, ..., θ_t^{33}, $\widetilde{V}_t^1$, ..., $\widetilde{V}_t^i$, ..., $\widetilde{V}_t^{33}$], where i is the index of a node. $\boldsymbol{z}$ is the set of the available measurements. $\boldsymbol{z}$ at the time is given by [P_t^1, $\widetilde{P}_t^2$, ..., $\widetilde{P}_t^i$, ..., $\widetilde{P}_t^{33}$, Q_t^1, $\widetilde{Q}_t^2$, ..., $\widetilde{Q}_t^i$, ..., $\widetilde{Q}_t^{33}$, V_t^1], where P_t^1, Q_t^1 and V_t^1 are the active power, reactive power and the voltage magnitude measured at the GSP, $\widetilde{P}_t^i$ and $\widetilde{Q}_t^i$ are the active power and reactive power at the i^{th} node forecasted by either a naive model or the load forecasting approach. We have also evaluated the performances of the DSE approach when additional real-time measurements are provided; these are the voltage magnitudes at some of the LV transformers (node 2-33). $h_i(\boldsymbol{x})$ in (3) is a measurement function relating $\boldsymbol{x}$ to a measurement z_i, so the residual $r_i = z_i - h_i(\boldsymbol{x})$ reflects the difference between the provided measurement and the measurement calculated from the estimated states. R is the variance of the measurements. It emphasizes the respective belief degrees of different measurements.

Newton's method is applied, in order to derive the optimal solution. It initializes all nodes' voltage magnitudes to 1 and the phasor angles to 0 and iteratively updates $\boldsymbol{x}$ through $\boldsymbol{x}^{k+1} = \Delta\boldsymbol{x}^{k+1} + \boldsymbol{x}^k$ in each iteration k, until $j(\boldsymbol{x})$ in (3) is minimized. The increment $\Delta\boldsymbol{x}^{k+1}$ is calculated through (4), where H is the measurement Jacobian, and $G(\boldsymbol{x}^k)$ is called the gain matrix which equals $H^T(\boldsymbol{x}^k)R^{-1}H(\boldsymbol{x}^k)$.

$$[G(\boldsymbol{x}^k)]\Delta\boldsymbol{x}^{k+1} = H^T(\boldsymbol{x}^k)R^{-1}[\boldsymbol{z} - h(\boldsymbol{x}^k)] \tag{4}$$

IV. Experiments

In this section we present the comparison between different LV transformer load forecasting when either the daily or the half-hourly collected SM data are used. The improvements when utilizing the load forecasting technique over the naive models with both data collection frequencies are assessed as well. Following this, the forecasted load based on the two data collection frequencies is used in DSE. The comparison of the respective DSE results shows the different suitability of the two SM data collection frequencies for power distribution system control.

A. Test Data

We use the Electricity Customer Behavior Trial Database from the Commission for Energy Regulation (CER) [17]. The database records the half hourly power consumptions between 14-07-2009 and 31-12-2010 for 6445 consumers, where 485 of them are small-medium enterprises, 1735 of them are other kind of consumers and the remaining 4225 participants are residential consumers. In this research we only use the records of the residential consumers which has the fewest missing data. 18 types of residential services had been provided to the consumers since 31-12-2009, because the trial was primarily designed for evaluating the impact of the smart meter installation on consumers electricity usage behaviour. We divide the consumers with the same service into a certain number of groups, which leads to 32 groups in total. We use each group to simulate the households downstream from an LV transformer in Fig. 1. We transform the consumption data to active power by multiplying the data with 2. The time series active power for each transformer is then achieved through aggregation.

B. Load Forecasting Evaluation

We use three models for load forecasting with both daily and half-hourly SM data collection: naive model, SSA-SVR and SSA-ANN. In total six cases of load forecasting are compared in terms of accuracy. The notations of the cases and the used approaches are listed as follows:

- Load forecasting on daily collected SM data
  - using *Naive Model 1*: *NM*$_1$
  - using *SSA-SVR*: *SSA-SVR*$_1$
  - using *SSA-ANN*: *SSA-ANN*$_1$
- Load forecasting on half-hourly collected SM data
  - using *Naive Model 2*: *NM*$_2$
  - using *SSA-SVR*: *SSA-SVR*$_2$
  - using *SSA-ANN*: *SSA-ANN*$_2$

The lower index of a notation denotes if this is for daily or half-hourly collected SM data (1 is daily, 2 is half-hourly).

Once the half-hourly active powers of the 32 transformers from 14-07-2009 to 31-12-2010 have been obtained, we use this data set to evaluate all six cases. The evaluations on the 32 transformers' data are carried out separately for each case. And the data of each transformer is divided into three portions for training, validating and testing purposes respectively. The validation data is used to select the load variables of attributes for the parallel models of *SSA-SVR*$_1$, *SSA-ANN*$_1$, *SSA-SVR*$_2$ and *SSA-ANN*$_2$. It is also used to determine the numbers of hidden neurons for the parallel models of *SSA-ANN*$_1$ and *SSA-ANN*$_2$.

The half-hourly active powers of each of the 32 transformers between 14-07-2010 and 31-12-2010 are forecasted for each of the cases. We use the absolute relative error as the criteria for evaluation. The absolute relative error of transformer b at time t is given by: $|(\widetilde{P}_t^b - P_t^b)/P_t^b|$, where $\widetilde{P}_t^b$ and P_t^b are the forecasted and actual active powers respectively. Furthermore, for each time point of the day (namely, 00:30, 01:00, ... and 24:00), we average the absolute relative errors over all the 171 days and the 32 transformers. The results are shown in Fig. 2.

The first three graphs in Fig. 2 show the comparisons in terms of error against the real state between the daily and half-hourly SM data collection frequencies when using the naive models, *SSA-SVR* or *SSA-ANN* for load forecasting respectively. The last graph shows the different performances of using ANN or linear-SVR as the learner of the load forecasting approach based on the half-hourly collected SM data. We can see from the first graph that the errors of *NM*$_2$ are smaller than the errors of *NM*$_1$ for most of time, in particular between

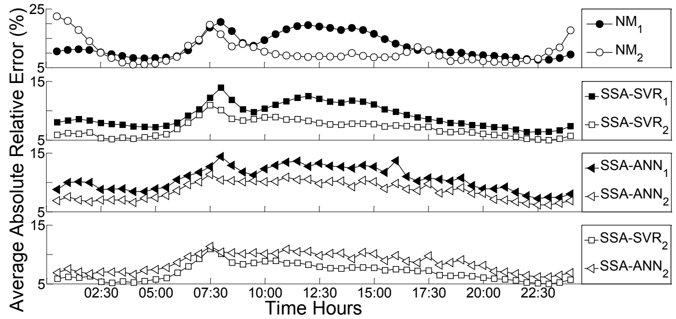

Fig. 2. Load forecasting performance of the six cases.

07:30 and 22:30, which is the typical demand active period for residential consumers. It is also clear from the second and third graphs that better performances are achieved when using the half-hourly collected SM data when load forecasting techniques are used rather than the naive models. The last graph shows higher load forecasting accuracies on the half-hourly collected SM data are achieved when using the linear-SVR as the learner over ANN.

Furthermore, we record the average rank (1 is best and 6 is worst) of a case across the 32 transformers for each time point of the day. We carry out a Friedman test which claims the average ranks are significant over time. It is then processed with the Bonferroni-Dunn post-hoc test ($p = 0.1$), which checks the significance against *SSA-SVR*$_1$ as the control learner. For intuitive presentation, we only report the results at 9 discrete time points. The significance test results are shown in Fig. 3.

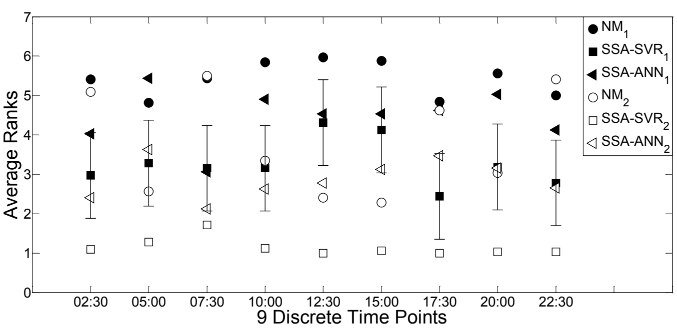

Fig. 3. Bonferroni-Dunn post-hoc test against *SSA-SVR*$_1$ as the controller. For each of the 9 discrete time points, the performance of a method is considered to be significantly different from *SSA-SVR*$_1$, if it is outside the vertical bar.

The average ranks of *SSA-SVR*$_1$ are significantly better than the ranks of *NM*$_1$ over all the time, and the pair of *SSA-SVR*$_2$ and *NM*$_2$ is the same. This means the utilization of the state of the art technique can significantly improve the load forecasting accuracies compared to naive models, independently if the data is collected once a day or half-hourly. Moreover, the ranks of *SSA-SVR*$_2$ are significantly better than the ranks of *SSA-SVR*$_1$, which also applies to *NM*$_2$ and *NM*$_1$ for most of the time when users' power consumption behaviours are more active. This means the forecasted loads with half-hourly collected SM data are usually more accurate, independently if the load forecasting algorithm is a state of the art technique or a simple similarity-based method.

C. Distribution State Estimation Evaluation

The 32 LV transformers' half-hourly active powers between 14-07-2010 and 31-12-2010 are obtained by aggregation. In order to obtain their reactive powers for the period, we set for each transformer a constant value to represent the ratio of reactive to active power $\frac{Q}{P}$. The same method has been used in [2] for the same purpose. Their time series voltages as well as the powers at the GSP (node 1) are calculated using power flow analysis. The data above is taken as the "actual states" of the system.

We evaluate the different performances when DSE is performed with *NM*$_1$, *NM*$_2$, *SSA-SVR*$_1$ and *SSA-SVR*$_2$ respectively. The active and reactive powers of the 32 transformers between 14-07-2010 and 31-12-2010 are forecasted for each case, and then used as the inputs to the DSE. In addition to the forecasted data, we assume the actual active and reactive powers and the voltages at the GSP are also available to DSE. DSE estimates the 32 transformers' time series voltage magnitudes. We show the relative errors of the estimated values for the four cases in Fig. 4. A point in the graph denotes the relative error of a transformer at a single time point. The relative error of the estimated voltage magnitude of transformer b at time t is given by: $(\widetilde{V}_t^b - V_t^b)/V_t^b$, where $\widetilde{V}_t^b$ and V_t^b are the estimated and actual voltage magnitudes respectively.

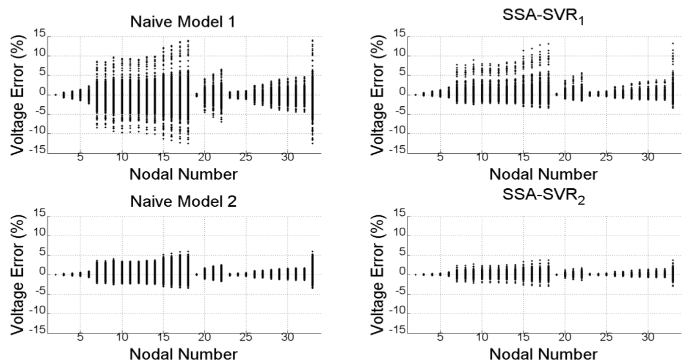

Fig. 4. Relative errors of the estimated voltage magnitudes (with real time data at GSP only).

DSE performs best with *SSA-SVR*$_2$. The maximum error for this case is under 4%. The maximum voltage error of *NM*$_2$ is slightly higher - just under 6%. As for the two cases handling the daily collected SM data, their maximum errors approach 13%. The results clearly show better performance (accuracy) of DSE when the SM readings are collected in every half an hour. However, even the 4% maximum error is too big to be used for reliable control decisions. In [2], [7] the threshold of 0.6% is considered as the maximum acceptable error. The most effective yet expensive way to improve the DSE performance is to introduce additional real-time measurements. Therefore, we assume the voltages at some of the LV transformers are observable. Further evaluations under this assumption are carried out by continuously increasing the number of real-time measurements. The following two figures show the DSE performances of the 4 cases with respectively 6 and 18 more real-time measurements points being available in power system MV nodes.

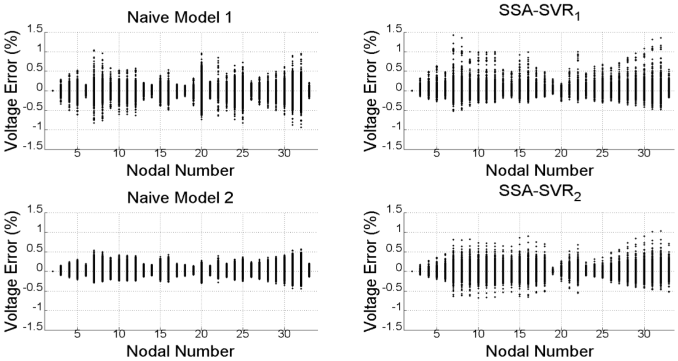

Fig. 5. Relative errors of the estimated voltage magnitudes (with real-time data at GSP and measurements of voltage magnitudes at 6 LV transformers).

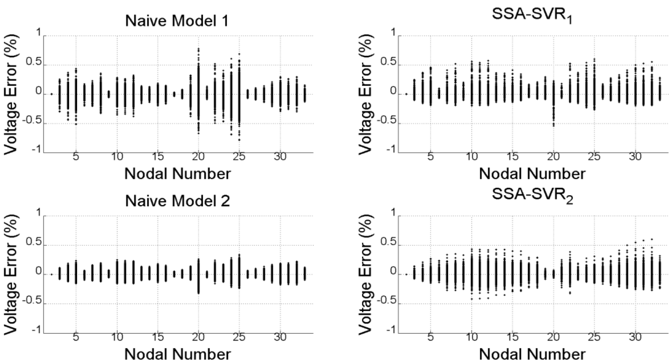

Fig. 6. Relative errors of the estimated voltage magnitudes (with real-time data at GSP and measurements of voltage magnitudes at 18 LV transformers).

In order to reduce the maximum voltage error to 0.6%, NM_2 needs 6 LV transformers to provide voltage measurements, while $SSA\text{-}SVR_2$ needs 15. The reason $SSA\text{-}SVR_2$ is not very helpful for DSE is because load forecasting techniques are not good at forecasting abnormal loads of special days like the Christmas holiday, and DSE is sensitive to big error. For the two cases based on the daily collected SM data, the maximum voltage error of $SSA\text{-}SVR_1$ firstly decreases to 0.6%, but it requires 18 real time measurements. The DSE results show neither data collection strategy alone is sufficient to estimate a system's status accurately - additional power system node real-time measurements are required. However significantly fewer additional measurements are required to achieve the same acceptable level of error if the data is collected half-hourly.

V. Conclusions

In this paper, we evaluated the suitability of the daily and half-hourly collected SM data for use in DSE. We first forecast the LV transformers' real-time loads using the delayed data. The forecasted loads are then fed into a DSE algorithm which estimates the transformers' voltages. Our significance test results show the accuracy of LV transformer load forecasting based on half-hourly collected SM data is significantly better independently if the utilized load forecasting algorithm is a state of the art technique or a simple (naive similarity) method. The DSE results further show that using the half-hourly collected SM data for power distribution system control needs much fewer real-time measurements inside the power system itself, thus less investment is required for additional measurement devices. Currently, the daily SM data collection strategy is primarily determined by considering only the billing function and user privacy. This research has developed a unique point of view (from a DSE perspective) that will hopefully help network operators to take a more system holistic view towards SM data value for power system operational use, and reflect this in requesting higher smart meter data collection frequency.

Acknowledgement

The authors wish to take this opportunity to thank the Commission for Energy Regulation (CER) for supplying the Electricity Customer Behavior Trial database. We would also like to express sincere thanks to Toshiba Research Europe Limited for funding this work.

References

[1] Z. Fan, Q. Chen, G. Kalogridis, S. Tan, and D. Kaleshi. The power of data: Data analytics for M2M and smart grid. In *IEEE ISGT Europe*, pages 1–8, Oct 2012.

[2] J. Wu et al. A robust state estimator for medium voltage distribution networks. *IEEE Trans. on Power Systems*, 28(2):1008–1016, 2013.

[3] Y. Chen et al. Short-term load forecasting: Similar day-based wavelet neural networks. *IEEE Transactions on Power Systems*, 25(1):322–330, Feb 2010.

[4] F.C. Schweppe and E.J. Handschin. Static state estimation in electric power systems. *Proceedings of the IEEE*, 62(7):972 – 982, July 1974.

[5] M. Baran and T.E. McDermott. Distribution system state estimation using ami data. In *IEEE PSCE*, 2009.

[6] Office of Gas and Electricity Markets. U.K. Smart metering implementation programme: Statement of design requirements, 2010.

[7] K. Samarakoon et al. Use of delayed smart meter measurements for distribution state estimation. In *Power and Energy Society General Meeting, IEEE*, pages 1–6, 2011.

[8] Eugene A. Feinberg and Dora Genethliou. Load forecasting. In Joe H. Chow, Felix F. Wu, and James Momoh, editors, *Applied Mathematics for Restructured Electric Power Systems*, Power Electronics and Power Systems, pages 269–285. Springer US, 2005.

[9] Z.H. Osman et al. Neural network based approach for short-term load forecasting. In *IEEE Power Systems Conference and Exposition*, 2009.

[10] M.B. Tasre et al. Hourly load forecasting using artificial neural network for a small area. In *ICAESM*, pages 379–385, March 2012.

[11] G. Darbellay and M. Slama. Forecasting the short-term demand for electricity: Do neural networks stand a better chance? *International Journal of Forecasting*, 16(1):71 – 83, 2000.

[12] E. Ceperic, V. Ceperic, and A. Baric. A strategy for short-term load forecasting by support vector regression machines. *IEEE Transactions on Power Systems*, 28(4):4356–4364, Nov 2013.

[13] S. Fan and L. Chen. Short-term load forecasting based on an adaptive hybrid method. *IEEE Transactions on Power Systems*, 21(1):392–401, Feb 2006.

[14] E.E. Elattar et al. Electric load forecasting based on locally weighted support vector regression. *IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 40(4):438–447, 2010.

[15] S. Fan et al. Comparative study on load forecasting technologies for different geographical distributed loads. In *2011 IEEE Power and Energy Society General Meeting*, pages 1–8, July 2011.

[16] M.E. Baran and F.F. Wu. Network reconfiguration in distribution systems for loss reduction and load balancing. *IEEE Transactions on Power Delivery*, 4(2):1401–1407, 1989.

[17] Commission for Energy Regulation (CER). Smart metering electricity customer behaviour trials (cbts). [online] http://www.ucd.ie/issda/data/commissionforenergyregulationcer/. Accessed: 08/01/2013.